Advanced Regression: 1c Random effects and hierarchical models (Part II)

Garyfallos Konstantinoudis

Epidemiology and Biostatistics, Imperial College London

21st February 2023

4. Random effects

1. Random effect model with random intercept

$$y_i = (\alpha_0 + u_k) + \beta x_i + \epsilon_i,$$

where $u_k \sim N(0, \sigma_u^2)$

2. Random effects model on both, the intercept and the slope

$$y_i = (\alpha_0 + u_k) + (\beta + w_k)x_i + \epsilon_i$$

where $w_k \sim N(0, \sigma_w^2)$

Group effects are random variables, also called random effects.

1. Random effect for the intercept $u_k \sim N(0, \sigma_u^2)$
2. Random effect for the slope $w_k \sim N(0, \sigma_w^2)$

Random intercept

1. Random effect model with random intercept

$$\begin{aligned} y_i &= (\alpha_0 + u_k) + \beta x_i + \epsilon_i, \\ &= \alpha_0 + \beta x_i + u_k + \epsilon_i, \end{aligned}$$

- Where α_0 is the intercept and β the regression coefficient.
- There are two distinct error terms
  1. Group-specific error
  $$u_k \sim N(0, \sigma_u^2)$$
  2. Individual-specific error
  $$\epsilon_i \sim N(0, \sigma^2)$$
- Note that u_k and ϵ_i are independent of each other.

Random effect model with random intercept

Interpretation of random intercept α_k:

$$\alpha_k = (\alpha_0 + u_k)$$

- α_0 is the global intercept
- u_k group-level variations around the global intercept

This is equivalent to assuming α_k is a **random variable** that follows a Normal distribution

$$\alpha_k \sim N(\alpha_0, \sigma_u^2)$$

Random effect model with random intercept

Multi-level interpretation (two levels of variability):

1. First level
   Defined on the individual level for observation $i = 1, ..., n$, similar to a standard linear regression

$$y_i = \alpha_k + \beta x_i + \epsilon_i,$$

2. Second level
   But the intercept is not fixed, it is a random variable

$$\alpha_k \sim N(\alpha_0, \sigma_u^2)$$

Random effect model with random intercept

Assumptions

- Slope of regression line is the same across all groups. Each group has a different intercept (α_k).
- But $\alpha_k \sim N(\alpha_0, \sigma_u^2)$ has now a common distribution which is estimated from **all observations**, and not just from the observations in a specific group as in fixed effects.
- We pool information across groups.

Consequences

- We control for group characteristics by including the group-specific intercept.
- Number of group-specific parameters to estimate is much smaller than in the fixed effect models (σ_u^2 vs k intercepts).

(Restricted) Maximum Likelihood estimation of random effect

$$y_i = \alpha_0 + \beta x_i + u_k + \epsilon_i,$$

Parameters to estimate are

- α_0, β intercept and regression coefficient
- σ_u^2, σ^2 variance components

Maximum Likelihood estimation is based on the Normal distribution of u_k and ϵ_i

- ML estimate for σ_u^2 requires subtracting 2 empirical estimates of variance $\rightarrow$ ML estimates for σ_u^2 can be negative.
- Restricted Maxium Likelihood (REML): Imposes positivity constraints on the variance estimates.

Random intercept in R

Implementations of Restricted Maximum Likelihood (REML) in R

- `lmer` function in the `lme4` package
- `lme` function in the `nlme` package

Focus here is the `lme` function in the `nlme` package.

`lme(fixed, data, random)`

- `fixed`: Formula $y \sim x$
- `random`: Formula $\sim 1 \mid \mathit{factor}$
- `data`: Dataset to use

R: Random intercept using lme

```
RandomIntercept = lme( chol ~ age, random = ~ 1 |
doctor, data = data.chol)
summary(RandomIntercept)
```

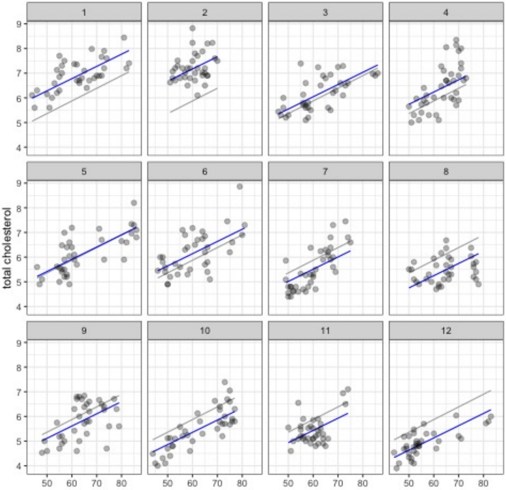
```
Linear mixed-effects model fit by REML
 Data: data.chol
      AIC      BIC    logLik
  828.697 845.035 -410.3485

Random effects:
 Formula: ~1 | doctor
        (Intercept)  Residual
StdDev:   0.6347908 0.5764246

Fixed effects: chol ~ age
                Value  Std.Error  DF  t-value p-value
(Intercept) 2.9060357 0.26477408 428 10.97553       0
age         0.0495831 0.00306279 428 16.18888       0
```

R: Random intercept using `lme`

```
RandomInterceptPredictions = fitted(RandomIntercept)
```



Random effect model and variance partition

Variance decomposition for observation i in group k

$$\begin{aligned} var(y_i) &= var(u_k + \epsilon_i) \\ &= var(u_k) + var(\epsilon_i) + 2cov(u_k, \epsilon_i) \\ &= \sigma_u^2 + \sigma^2 + 0 \end{aligned}$$

Further we can look at the covariance of observations

- i and i' within group k

$$cov(y_i, y_{i'}) = cov(u_k + \epsilon_i, u_k + \epsilon_{i'}) = \sigma_u^2$$

- i and i' from different groups k and k'

$$cov(y_i, y_{i'}) = cov(u_k + \epsilon_i, u_{k'} + \epsilon_{i'}) = 0$$

Random effect model and variance partition

Variability between and within groups

Intra-class correlation coefficient ρ

$$\rho = cor(y_i, y_{i'}) = \frac{cov(y_i, y_{i'})}{\sqrt{var(y_i)var(y_{i'})}} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}$$

Interpretation:

- Intra-class correlation coefficient ρ is the correlation between two observations i and i' in the same group.
- It is the ratio of between-group variance σ_u^2 over the total variance.
- If $\rho \to 0$ there is little variation explained by the grouping and we might consider a model without the random effect.
- Any restrictions?

Variance partition in R

```
summary(RandomIntercept)
```

```
Random effects:
 Formula: ~1 | doctor
        (Intercept)  Residual
StdDev:   0.6347908 0.5764246
```

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2} = \frac{0.6347908^2}{0.6347908^2 + 0.5764246^2} \approx 0.54$$

Interpretation:

- There is substantial evidence for between-group heterogeneity.
- More than half of the total variance can be explained by the between-group variance.
- It is beneficial to include the random effects on the intercept.

Random effect model with random intercept and random slope

1. Random effects model on both, the intercept and the slope

$$y_i = (\alpha_0 + u_k) + (\beta + w_k)x_i + \epsilon_i$$

- There are three distinct error terms
  1. Group-specific error of the intercept

$$u_k \sim N(0, \sigma_u^2)$$

  2. Group-specific error of the regression slope

$$w_k \sim N(0, \sigma_w^2)$$

  3. Individual-specific error

$$\epsilon_i \sim N(0, \sigma^2)$$

- Note that u_k and w_k are correlated and independent of ϵ_i.

Random effect model with random intercept and random slope

Assumptions

- Each group has a different intercept ($\alpha_k = \alpha_0 + u_k$) and a different regression slope ($\beta_k = \beta + w_k$).
- We allow for correlation between α_k and β_k.
- Both, $\alpha_k \sim N(\alpha_0, \sigma_u^2)$ and $\beta_k \sim N(\beta, \sigma_w^2)$ have a common distribution which is estimated from **all observations**, and not just from the observations in a given group as in fixed effects.
- We pool information across groups.

Consequences

- Including a random slope can be interpreted as creating an interaction between the group and the strength of association.
- We only have three additional parameters in the model: σ_u^2, σ_w^2 and $cor(\sigma_u, \sigma_w)$.

R: Random intercept and slope using lme

```
RandomSlope = lme( chol ~ age, random = ~ 1+age |
doctor, data = data.chol)
summary(RandomSlope)
```

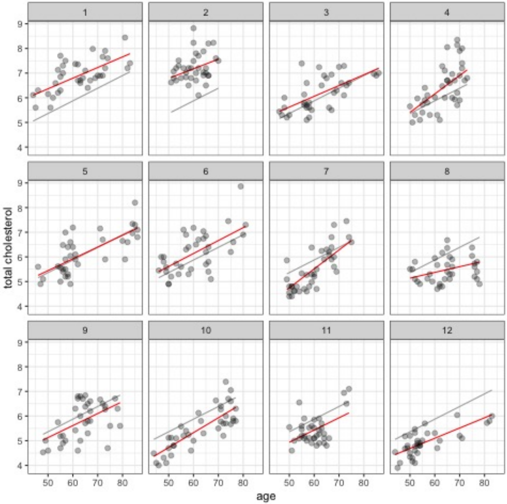
```
Linear mixed-effects model fit by REML
 Data: data.chol
      AIC      BIC    logLik
 821.9886 846.4956 -404.9943

Random effects:
 Formula: ~1 + age | doctor
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr
(Intercept) 1.28163791 (Intr)
age         0.01771585 -0.872
Residual    0.55997509

Fixed effects: chol ~ age
                Value Std.Error  DF  t-value p-value
(Intercept) 2.8791744 0.4215200 428 6.830458       0
age         0.0500704 0.0060597 428 8.262837       0
```

R: Random intercept and slope using lmer

```
RandomSlopePredictions = fitted(RandomSlope)
```



Variables on individual level and group level

When considering variables or predictors we need to distinguish:

- Individual-level variables
- Group-level variables, that are the same for all observations in a group

GP example:

- Individual-level variables: Age and sex of patient
- Group-level variables: Age of doctor

```
   chol doctor age   bmi agedoc sex
1 7.13       1  54 27.39     55   0
2 7.70       1  55 29.10     55   0
3 7.30       1  56 27.90     55   0
4 6.89       1  71 26.67     55   1
5 6.90       1  72 26.70     55   1
6 7.90       1  73 29.70     55   1
```

Variables on individual level and group level

$$y_i = (\alpha_0 + u_k) + (\beta + w_k)x_i + \gamma x_g + \epsilon_i$$

Example: GP data

```
RandomCov = lme( chol ~ age + agedoc, random = ~
1+age | doctor, data = data.chol)
summary(RandomCov)
```

```
Fixed effects: chol ~ age + agedoc
                 Value Std.Error  DF   t-value p-value
(Intercept) -2.7897788 1.1824050 428 -2.359411  0.0188
age          0.0501492 0.0060673 428  8.265423  0.0000
agedoc       0.1280030 0.0253576  10  5.047908  0.0005
```

Model comparison

- Likelihood-ratio test for nested models:
  - Models must have the same fixed effects. Does not work with group-level covariates.
  - Model with smaller - log likelihood is better (better model fit).
- Akaike information criterion (AIC)
  - Model with the smaller AIC is better (less information loss).

GP example:

◇ Model A (Random intercept)

```
modelA = lme( chol ~ age, random = ~1 | doctor, data = data.chol)
```

◇ Model B (Random intercept and slope)

```
modelB = lme( chol ~ age, random = ~ 1+age | doctor, data = data.chol)
```

◇ Model C (Random intercept and slope and group covariate)

```
modelC = lme( chol ~ age + agedoc, random = ~ 1+age | doctor, data = data.chol)
```

Model comparison

- Likelihood-ratio test for nested models
  (Model A is nested in Model B)
  `anova(modelA,modelB)`

```
> anova(modelA, modelB)
       Model df      AIC      BIC    logLik   Test L.Ratio p-value
modelA     1  4 828.6970 845.0350 -410.3485
modelB     2  6 821.9886 846.4956 -404.9943 1 vs 2 10.7084  0.0047
```

- AIC for non-nested models
  `anova(modelB,modelC)`

```
> anova(modelB, modelC)
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
modelB     1  6 821.9886 846.4956 -404.9943
modelC     2  7 815.6956 844.2712 -400.8478 1 vs 2 8.292926   0.004
Warning message:
In anova.lme(modelB, modelC) :
  fitted objects with different fixed effects. REML comparisons are not
meaningful.
```

Generalised linear mixed models

- Generalised Linear Mixed models (GLMM) can be used to adapt linear mixed models to outcomes that do not follow a Normal distribution.
- The package `lme4` includes the function `glmer` that can fit GLMMs.

```
glmer(formula, family = gaussian)
```

Formula:

- `y~x` to specify outcome and predictors
- `+ (1 | factor)` add random intercept depending on factor
- `+ x + (x | factor)` add random slope depending on factor

Take away: Fixed and random effect

- Fixed effect models can account for group structure but many parameters need to be estimated and no information is shared between groups.
- Random effect models treat group-specific parameters as random variables.
- Instead of estimating one parameter for each group, random effect models only estimate the distribution parameter of the random variable.
- Thus, they pool information across groups.
- The intra-class coefficient gives a measure of how relevant the group structure is.
- Implementation in `R`: `lme()` function in the `nlme` package.
- Models including both, fixed and random effects, are often called linear mixed models.